CALIBRATION IS GROUPING: VR-SAG WITH INTRA-GROUP VARIANCE CONTROL AND LOGIT-CLUSTER EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate click-through and conversion-rate estimates are pivotal for bid optimization in large-scale advertising, yet modern deep CTR/CVR models are often miscalibrated. Classical global calibrators (Platt scaling, isotonic regression) and feature-based binning struggle to capture latent user-item heterogeneity. We approach calibration through the lens of *latent, calibration-aware groupings* and propose **Variance-Reduced Semantic-Aware Grouping (VR-SAG)**, a lightweight post-hoc layer over a frozen backbone that (i) learns calibration-aware groups in embedding space, (ii) fits per-group temperature+bias calibrators, and (iii) explicitly penalizes intra-group variance to tighten probability spreads. A group-wise decomposition of the Brier score guides our design. It separates within-group variance from score–label covariance and motivates a grouping regularizer. To decouple evaluation from training, we introduce **Logit-Cluster Calibration Error (LCCE)**, an unsupervised fixed-partition metric obtained via K-means in logit space; LCCE aligns with the reliability term of proper scores while avoiding pitfalls of trainable grouping heads used as metrics. Across offline CTR logs and **AdAuction** (an ad-auction dataset with simulator-defined oracle CTRs), VR-SAG improves the reported calibration metrics over the corresponding SAG variants. Its additional computation scales linearly with the number of grouping heads and groups.

1 INTRODUCTION

Machine learning recommender systems underpin virtually every modern advertising platform, orchestrating the selection and pricing of *tens of billions* of ad impressions each day (Covington et al., 2016; Zhang et al., 2014b). For each impression, the model reports two probabilities—click-through rate (CTR) and conversion rate (CVR)—whose precision is crucial for both platform revenue and advertiser return (Richardson et al., 2007; He et al., 2014). Because an auction bid equals an advertiser's private value times one of these predicted probabilities, even modest calibration errors propagate into mispriced traffic and distorted budget pacing (McMahan et al., 2013). To satisfy strict latency and scale requirements, production systems typically employ deep architectures such as Wide & Deep (Cheng et al., 2016) and DeepFM (Guo et al., 2017b). While we motivate both CTR and CVR at a high level, all experiments in this paper focus on impression-level CTR calibration over exposure logs; extending the same ideas to CVR under stronger selection bias is left as future work.

Despite strong ranking performance, these models often produce *miscalibrated* probabilities: after grouping predictions into narrow bins, the observed click frequency rarely matches the average score. Such miscalibration erodes auction efficiency and reduces revenue (Lin et al., 2024), motivating extensive work on calibration for ads ranking (McMahan & Muralidharan, 2012; Fan et al., 2023; Borisov et al., 2018; Chaudhuri et al., 2017; Sheng et al., 2023). Existing approaches either learn a single global mapping (e.g., Platt scaling (Platt et al., 1999), isotonic regression (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005), temperature scaling (Guo et al., 2017a)), which can leave residual error within subpopulations, or rely on predefined metadata partitions (multicalibration (Hébert-Johnson et al., 2018), field-aware methods (Pan et al., 2020; Yang et al., 2024)), whose chosen fields may leave residual heterogeneity and opposite biases within a group.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these gaps, we adapt Semantic-Aware Grouping (SAG) (Yang et al., 2023) for CTR/CVR calibration and introduce three contributions. First, we derive a group-wise Brier-loss decomposition that separates variance and covariance contributions, and propose *Variance-Reduced SAG* (VR-SAG), which jointly learns per-group temperatures and biases while penalizing intra-group variance to enforce tighter, more coherent partitions. Second, we decouple evaluation from the trainable grouping head by defining the *Logit-Cluster Calibration Error* (LCCE), an unsupervised, fixed-partition metric in logit space that aligns with the reliability term of proper scoring rules. Finally, we construct **AdAuction**, a large-scale ad-auction dataset with oracle CTRs generated by an internal simulator that mirrors industrial auction workflows. In summary:

- We use a group-wise Brier-loss decomposition to motivate VR-SAG, which adds intra-group variance regularization to the SAG architecture.
- We propose LCCE, a fixed logit-space grouping metric that audits group-mean calibration independently of the learned grouping head.
- We construct **AdAuction**, a simulated ad-auction dataset with oracle CTRs generated by an internal simulator, facilitating rigorous and reproducible benchmarking of calibration techniques in large-scale recommender systems.

2 METHOD

We begin by reviewing binary-CTR prediction and the Expected Calibration Error (Sec. 2.1), then introduce Semantic-Aware Grouping (SAG), which applies group-specific temperatures over a frozen backbone (Sec. 2.2). Building on SAG, we derive a group-wise Brier-loss decomposition that separates variance and covariance contributions (Sec. 2.3) and propose variance-regularized VR-SAG to address it (Sec. 2.4). Finally, we present the Logit-Cluster Calibration Error (LCCE), an unsupervised, fixed-partition metric for calibration evaluation (Sec. 2.5).

2.1 BACKGROUND

Let an impression be represented by a feature vector $x \in \mathbb{R}^d$ obtained by concatenating user descriptors, ad metadata and real-time context, and let the click indicator be $y \in \{0, 1\}$ ($y = 1$ means the user clicked the ad). A predictor $f_\theta : \mathbb{R}^d \to [0, 1]$ parameterized by θ outputs the raw click-through probability $\hat{p} = f_\theta(x)$. Its penultimate layer produces a hidden representation $z(x) \in \mathbb{R}^m$, and its last linear layer returns a single logit $o(x) \in \mathbb{R}$ before the final sigmoid activation. The network is trained by minimizing the average negative log-likelihood

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big], \tag{1}$$

a proper scoring rule (Gneiting & Raftery, 2007). Its population optimum recovers the conditional click probability; approximate optimization over a restricted model family does not by itself guarantee calibration (Błasiok et al., 2023).

A predictor is *well-calibrated* when the conditional click frequency equals its score, i.e. $\Pr(Y = 1 \mid \tilde{p} = q) = q$ for all $q \in [0, 1]$. The population absolute calibration error is $\mathbb{E}[|\mathbb{E}[Y \mid \hat{p}] - \hat{p}|]$. Practitioners approximate this quantity with the Expected Calibration Error (ECE), the weighted average gap between predicted probability and empirical click rate across probability bins. Formally, partition $[0, 1]$ into M equal-width bins $B_1, \ldots, B_M$; then

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \underbrace{\tfrac{1}{|B_m|} \sum_{i \in B_m} y_i}_{\text{acc}(B_m)} - \underbrace{\tfrac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i}_{\text{conf}(B_m)} \Big|, \tag{2}$$

with lower values indicating better alignment between predicted probabilities and observed outcomes.

2.2 SEMANTIC-AWARE GROUPING FOR CTR CALIBRATION

Semantic-Aware Grouping (SAG)(Yang et al., 2023) augments the frozen backbone with a lightweight grouping head. A weight matrix $W \in \mathbb{R}^{m\times K}$ and bias $b \in \mathbb{R}^K$ transform the embedding $z(x)$ into soft group weights

$$q_k(x) = \mathrm{softmax}\big(z(x)^\top W + b\big)_k, \quad k = 1, \dots, K, \tag{3}$$

where K is the chosen number of latent regions in embedding space that are expected to share similar calibration behaviour, rather than necessarily corresponding to single human-interpretable fields. We denote this softmax-based grouping head by g_ϕ, with parameters $\phi = (W, b)$, so that $g_\phi(x) = q(x) = (q_1(x), \dots, q_K(x))$. For each group we keep a single temperature $\tau_k > 0$. The calibrated click-through probability (binary) is then the mixture

$$\tilde{p}(x) = \sum_{k=1}^{K} q_k(x)\ \sigma\big(o(x)/\tau_k\big), \tag{4}$$

where $\sigma(t) = (1 + e^{-t})^{-1}$.

All added parameters $\phi = (W, b)$ and $\{\tau_k\}_{k=1}^K$ are learned jointly on a held-out validation set D_{val} with the SAG objective

$$\mathcal{L}_{\text{SAG}} = -\frac{1}{|D_{\text{val}}|} \sum_{(x,y)\in D_{\text{val}}} \log\Big[\sum_{k=1}^{K} q_k(x)\ \big(y\,\tilde{p}_k(x) + (1-y)\big(1-\tilde{p}_k(x)\big)\big)\Big] + \lambda\|W\|_2^2, \tag{5}$$

where $\lambda > 0$ regularizes the grouping weights, and $\tilde{p}_k(x) = \sigma\big(o(x)/\tau_k\big)$. We retain the soft weights $q_k(x)$ during both calibration and serving, avoiding hard arg-max reassignment. At inference the extra cost is one $m \times K$ matrix-vector product and K scalar operations, with cost $O(mK)$ for one grouping head.

Why grouping can help. Soft assignments let an impression share several group-specific corrections. The groups represent latent regions of the backbone embedding; their correspondence to user cohorts or ad attributes requires separate analysis. In our usage, "semantic" therefore refers to calibration-aware structure already encoded in $z(x)$ by the CTR/CVR backbone, not to perfectly disentangled or manually labelled concepts.

2.3 DECOMPOSITION OF PROPER SCORING RULES WITH SEMANTIC GROUPS

Probabilistic models are ideally evaluated using *proper scoring rules* (Gneiting & Raftery, 2007), i.e., loss functions whose expected value is uniquely minimized when the predicted probabilities coincide with the true data-generating probabilities. In our binary CTR setting, both negative log-likelihood and the Brier score are proper, which motivates using them as the main training and analysis tools in what follows. Popular calibration metrics such as ECE, while intuitive, lack this property and can be gamed without improving true predictive fidelity. To bridge this gap we expose how calibration terms reappear inside a proper scoring rule once predictions are partitioned into semantic groups. For clarity we detail the case of the ***Brier score***; the same reasoning carries over to other proper scoring rules—including cross-entropy— yielding analogous insights with different algebraic constants[1]. The decomposition that follows clarifies when reducing a calibration error genuinely lowers a proper loss and when it merely provides a misleading signal.

For soft assignments, interpret conditioning as $\mathbb{E}[Z \mid G_k] = \mathbb{E}[q_k(X)Z]/\mathbb{E}[q_k(X)]$ and $w_k = \mathbb{E}[q_k(X)]$. Hard groups are the indicator-weight special case (Appendix A.2). Let G_k denote the kth group and write

$$w_k = \Pr(G_k), \quad \pi_k = \Pr(Y = 1 \mid G_k), \quad \mu_k = \mathbb{E}[\hat{p} \mid G_k],$$

where $\hat{p} = f_\theta(x)$ is the *uncalibrated* probability output of the frozen backbone. Write the within–group variance $\sigma_k^2 = \mathrm{Var}(\hat{p} \mid G_k)$ and covariance $\gamma_k = \mathrm{Cov}(\hat{p}, Y \mid G_k)$.

[1] See the Appendix for detailed proofs and analysis of other proper scoring rules.

For binary events the Brier loss is $S(Y, \hat{p}) = (Y - \hat{p})^2$. Conditioning on G_k and using $Y^2 = Y$ gives

$$\mathbb{E}[S \mid G_k] = \pi_k - 2\,\mathbb{E}[\hat{p}\,Y \mid G_k] + \mathbb{E}[\hat{p}^2 \mid G_k].$$

Because $\mathbb{E}[\hat{p}\,Y \mid G_k] = \mu_k \pi_k + \gamma_k$ and $\mathbb{E}[\hat{p}^2 \mid G_k] = \mu_k^2 + \sigma_k^2$, we obtain

$$\mathbb{E}\big[(Y - \hat{p})^2 \mid G_k\big] = \pi_k(1 - \pi_k) + (\pi_k - \mu_k)^2 + \sigma_k^2 - 2\gamma_k.$$

Averaging over groups yields the **grouping decomposition**

$$\mathbb{E}[S] = \underbrace{\bar{Y}\,(1 - \bar{Y})}_{\text{UNC}} + \underbrace{\sum_k w_k(\mu_k - \pi_k)^2}_{\text{REL}} - \underbrace{\mathrm{Var}(\pi_k)}_{\text{RES}} + \underbrace{\sum_k w_k(\sigma_k^2 - 2\gamma_k)}_{\Delta}, \tag{6}$$

where $\bar{Y} = \mathbb{E}[Y]$ denotes the marginal click rate. Using the law of total variance, $\sum_k w_k \pi_k(1 - \pi_k) = \bar{Y}(1 - \bar{Y}) - \mathrm{Var}(\pi_k)$. The classical UNC + REL − RES(Murphy, 1973) form is recovered whenever $\Delta = 0$. Constant forecasts within every group, with $\sigma_k^2 = \gamma_k = 0$, give a sufficient condition; cancellation can also yield $\Delta = 0$.

The correction Δ can have either sign. Cauchy–Schwarz gives $|\gamma_k| \le \sigma_k \sqrt{\pi_k(1 - \pi_k)}$, but a small click rate alone does not imply variance dominance. Learning a partition changes its masses, means and covariances together. In particular, changing the partition while holding predictions fixed leaves the total Brier score unchanged. The decomposition therefore motivates variance control as an inductive bias for learning groups; it does not prove that the regularizer decreases Brier loss or guarantees calibration.

Definition 2.1 (Grouping Calibration Error (GCE))**.** Given the latent regions $\{G_k\}_{k=1}^K$ induced by the grouping function g_ϕ, let $w_k = \Pr(G_k)$, $\mu_k = \mathbb{E}[\hat{p} \mid G_k]$ and $\pi_k = \Pr(Y = 1 \mid G_k)$. The *Grouping Calibration Error* of a probabilistic predictor f_θ with respect to this partition is

$$\mathrm{GCE}(g_\phi; f_\theta) \;=\; \sum_{k=1}^{K} w_k \big(\mu_k - \pi_k\big)^2. \tag{7}$$

Equation 7 is identical to the **REL** term in the grouping decomposition of the Brier loss given in equation 6. The squared gap matches the Brier reliability component; standard ECE instead uses an absolute gap. GCE measures agreement of group means, which is weaker than calibration at every score within each group. Lower GCE alone does not imply lower Brier loss because the other decomposition terms can change. VR-SAG trains with NLL and a variance penalty; GCE serves as an analytical diagnostic.

Re-expressing classical calibration metrics via grouping. The grouping perspective unifies several existing metrics:

- **Singleton groups.** When every impression forms its own group ($K = n$ and $G_k = \{(x_i, y_i)\}$), we have $w_k = \frac{1}{n}$, $\mu_k = \hat{p}_i$, and $\pi_k = y_i$. Substituting these quantities in equation 7 gives $\mathrm{GCE} = \frac{1}{n}\sum_{i=1}^n (\hat{p}_i - y_i)^2$, *exactly the Brier score*.[2]
- **Probability-based binning.** If instances are grouped according to their predicted probability—for example into M equal-width or equal-frequency bins—each bin B_m acts as a region G_k. Then μ_k equals the bin's average confidence, π_k equals its empirical accuracy, and GCE reduces to the weighted sum of squared (accuracy–confidence) gaps that underlies the squared-ECE variant.

These examples show that the partition determines which discrepancies a grouping metric exposes. A fixed coarse partition can hide errors that cancel within a group. We therefore report grouping metrics together with proper losses.

[2]With singleton groups the uncertainty and resolution terms in equation 6 cancel, while $\Delta = 0$, so the empirical Brier score coincides with the reliability component. This finite-sample identity does not identify an individual impression's true click probability.

2.4 VARIANCE-REDUCED SEMANTIC-AWARE GROUPING (VR-SAG)

Let the validation set contain $n = |D_{\text{val}}|$ impressions indexed by $i = 1, \dots, n$. For each impression x_i the frozen backbone produces a raw score $\hat{p}_i = f_\theta(x_i)$, a hidden vector $z_i = z(x_i)$ and logit $o_i = o(x_i)$. The *grouping head* g_ϕ (parameters $\phi = (W, b)$) returns soft assignments

$$q_{ik} \;=\; \mathrm{softmax}(z_i^\top W + b)_k \quad k = 1, \dots, K,$$

to K latent regions $\{G_k\}$. For each group we keep a temperature $\tau_k > 0$ and a bias $\beta_k \in \mathbb{R}$, so that the calibrated probability is

$$\tilde{p}_i \;=\; \sum_{k=1}^{K} q_{ik}\, \sigma\big(o_i/\tau_k + \beta_k\big). \tag{8}$$

This amounts to per-group Platt scaling on the backbone logits, mixed by the soft semantic assignments q_{ik}. When all $\beta_k = 0$ this reduces to temperature scaling; learning both $\{\tau_k\}$ and $\{\beta_k\}$ recovers per-group Platt scaling.

We define three empirical statistics per group:

$$\hat{w}_k \;=\; \frac{1}{n}\sum_{i=1}^{n} q_{ik}, \quad \hat{\mu}_k \;=\; \frac{1}{n\hat{w}_k}\sum_{i=1}^{n} q_{ik}\,\hat{p}_i, \quad \hat{\sigma}_k^2 \;=\; \frac{1}{n\hat{w}_k}\sum_{i=1}^{n} q_{ik}\,\big(\hat{p}_i - \hat{\mu}_k\big)^2, \tag{9}$$

i.e., the *normalized soft mass*, the mean uncalibrated score, and the within-group variance, respectively. The normalization defines weighted empirical moments for the optimization objective, not an unbiased finite-sample variance estimator. For a fixed grouping function these moments are consistent under i.i.d. sampling.

Variance reduction. We regularize the spread of frozen backbone probabilities within each learned group. This shapes the assignments while the likelihood term fits the calibrated outputs:

$$\mathcal{L}_{\text{VAR}} \;=\; \lambda_v \sum_{k=1}^{K} \hat{w}_k\, \hat{\sigma}_k^2, \tag{10}$$

with tunable weight $\lambda_v \geq 0$. The regularizer acts on $\hat{p}$, whereas the NLL acts on $\tilde{p}$. Thus it directly trains the grouping head, without directly forcing the calibrated outputs toward a constant. Applying the Brier identity to $\tilde{p}$ requires recomputing its own means, variances and covariances. Equation equation 10 is a proxy for grouping, not the calibrated predictor's heterogeneity term.

VR-SAG objective. Combining these elements, the validation-time loss is

$$\begin{aligned}\mathcal{L}_{\text{VR-SAG}} &= -\frac{1}{n}\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} q_{ik}\,\big(y_i\,\tilde{p}_{ik} + (1-y_i)\big(1-\tilde{p}_{ik}\big)\big)\Big] + \lambda\,\|W\|_2^2 + \mathcal{L}_{\text{VAR}} \\ &= \mathcal{L}_{\text{SAG-B}} + \lambda_v \sum_{k=1}^{K} \hat{w}_k\,\hat{\sigma}_k^2,\end{aligned} \tag{11}$$

where $\tilde{p}_{ik} = \sigma(o_i/\tau_k + \beta_k)$ and $\mathcal{L}_{\text{SAG-B}}$ denotes the SAG objective in equation 5 extended to include per-group biases $\{\beta_k\}$.

The NLL fits the final probabilities, and the variance penalty favors groups with similar backbone scores. Their joint objective does not guarantee distinct semantic groups or exclude score-based grouping. With $\lambda_v = 0$, the stated objective recovers the corresponding SAG objective with the same bias configuration.

Positive temperatures make each component monotone in its input logit. Since $q_k(x)$ varies across impressions, the mixture need not preserve their global ranking. AUC therefore requires separate measurement (Appendix G).

For one linear grouping head, W, b, τ and β contain $K(m+3)$ parameters in the PS variant; the TS variant has $K(m+2)$. With P grouping heads, parameter storage and inference work scale by P, giving $O(PmK)$ work. The experiments use $P = 10$ learned partitions with $K = 3$ groups each. We distinguish these learned partitions from the evaluation clusters below and from repeated test-set splits. Pipeline timing appears in Appendix D; it does not measure serving latency.

2.5 LOGIT-CLUSTER CALIBRATION ERROR (LCCE)

As introduced in Sec. 2.3, the Grouping Calibration Error in equation 7 measures the reliability term of a proper scoring rule under a partition g_ϕ. While GCE benefits from data-adaptive partitions, its coupling to the trainable grouping head can mask true miscalibration by driving GCE down even when predictions remain poorly aligned with outcomes. To preserve the low-variance, model-aware slicing of GCE *without* a learned component, we define the *Logit-Cluster Calibration Error* by applying the same squared-gap measure to clusters formed in logit space.

Let r_i denote the probabilities of the predictor under evaluation: $r_i = \hat{p}_i$ for the backbone and $r_i = \tilde{p}_i$ for a calibrated model. Fit K_c-means to the frozen backbone logits $\{o_i\}$, using no labels, and keep the resulting clusters $M_1, \ldots, M_{K_c}$ fixed across methods on the same evaluation sample. Define

$$w_j = \frac{|M_j|}{n}, \qquad \mu_j(r) = \frac{1}{|M_j|} \sum_{i \in M_j} r_i, \qquad \pi_j = \frac{1}{|M_j|} \sum_{i \in M_j} y_i.$$

Then

$$\mathrm{LCCE}_{K_c}(r) = \sum_{j=1}^{K_c} w_j \big(\mu_j(r) - \pi_j\big)^2. \tag{12}$$

This is GCE on the fixed logit partition. We use $K_c = 4$ evaluation clusters, distinct from the $K = 3$ learned groups, and average over four clustering seeds. Each seed uses the same clusters for every compared calibrator. Fixing the clusters removes optimization of the evaluation partition, but does not prevent within-cluster cancellation or guarantee full conditional calibration.

Why logits? One-dimensional K_c-means yields intervals with boundaries between adjacent centroids. These intervals adapt to logit density and generally have unequal widths and unequal sample counts. The sigmoid maps them to nonuniform probability intervals. This construction needs neither labels nor the trainable grouping head, but retains sensitivity to cluster count and initialization. Appendix C reports the available sensitivity analyses.

3 EXPERIMENT

We analyze calibration metrics and evaluate our method offline on AliCCP (Ma et al., 2018), AliExpress (AE), and AdAuction.

Dataset with ground-truth CTR We evaluate on **AdAuction**, an impression ad-auction dataset with oracle CTRs generated by an internal simulator. Like AuctionNet (Su et al., 2024), the underlying simulator retains the core workflow logic of industrial advertising systems—where auto-bidding agents process advertiser objectives, execute bid decisions, and collect post-auction feedback—thereby simulating inherent challenges such as sample selection bias (SSB). Specifically, its bidding mechanism mimics the natural overestimation issue: an overestimated ad item tends to win auctions more frequently and gains higher exposure during ranking, reflecting real-world biases in ad delivery.

Unlike AuctionNet, which solely records observable metrics (e.g., clicks/conversions), AdAuction incorporates ground-truth click-through rates as synthetic labels in its exposure data—an oracle signal inaccessible in real-world applications. This design enables direct calibration-error measurement against known truth values, a critical advantage for validating probabilistic prediction models that remains fundamentally unattainable in operational advertising platforms. The dataset contains 15M exposure samples with 451K clicks, and we plan to release the AdAuction feature schema and labels upon acceptance.

3.1 EXPERIMENTAL COMPARISON

We randomly partition a validation set D_{val} from the standard training set (10% for validation), and reserve 10% of the standard test set as a hold-out calibration set D_{ho}. For each dataset–model combination, we perform 100 random test-set splits and report the average performance over 100 trials for each method. The original analysis used paired t-tests over these splits. Repeated splits

Table 1: Comparison of calibration methods (reported mean values). Here oracle MAE is $\mathbb{E}[|\pi - r|]$ and oracle RMSE is $\sqrt{\mathbb{E}[(\pi - r)^2]}$, where r is the evaluated predictor. The RMSE label is provisional: the archived routine takes a square root, but its correspondence to the final table requires verification. The numerical values remain unchanged. *Note:* On **AdAuction**, oracle MAE and RMSE use simulator CTR π; on **AliCCP** and **AE**, we report ECE/LCCE.

	AdAuction				AliCCP		AE	
Method	ECE	LCCE	MAE	RMSE	ECE	LCCE	ECE	LCCE
Uncal	0.0339	0.0342	0.0352	0.0527	0.2131	0.2161	0.2562	0.2562
Histogram binning	0.0049	0.0113	0.0170	0.0325	0.0185	0.0210	0.0168	0.0222
Isotonic regression	0.0056	0.0090	0.0123	0.0282	0.0076	0.0081	0.0079	0.0094
Platt scaling	0.0063	0.0124	0.0157	0.0297	0.0056	0.0071	0.0065	0.0077
Temperature scaling	0.0062	0.0124	0.0158	0.0296	0.0056	0.0059	0.0052	0.0061
SAG+PS	0.0058	0.0100	0.0131	0.0247	0.0027	0.0031	0.0027	0.0036
SAG+TS	0.0054	0.0099	0.0130	0.0249	0.0013	0.0017	0.0016	0.0020
VR-SAG+PS(ours)	0.0054	0.0083	0.0118	0.0237	0.0008	0.0008	0.0008	0.0010
VR-SAG+TS(ours)	0.0053	0.0083	0.0117	0.0239	0.0003	0.0005	0.0003	0.0005

can overlap, so they do not represent 100 independent datasets; we interpret the reported means descriptively. Hyperparameters of comparative methods are tuned following their original papers, using 5-fold cross-validation. Unless otherwise noted, we fix the number of groups to 3 and the number of partitions to 10. The regularization strength is set to $\lambda = 0.1$, and the variance-penalty coefficient to $\lambda_v = 0.5$, following a similar tuning protocol as the baselines.

We compare the uncalibrated backbone (Uncal) against standard post-hoc calibrators and grouping-based methods: Histogram binning (Zadrozny & Elkan, 2001) (bin-wise averaging), Isotonic regression(Zadrozny & Elkan, 2002) (monotone piecewise mapping), Platt scaling (Platt et al., 1999) (logistic bias+scale), Temperature scaling (Guo et al., 2017a) (single temperature), SAG (Yang et al., 2023) (embedding-based semantic groups with per-group temperatures), and VR-SAG (ours) (SAG with intra-group variance control, evaluated with both PS/TS).

Table 1 reports lower ECE and LCCE for VR-SAG than for the corresponding SAG variants on all three datasets. Histogram binning has the lowest AdAuction ECE (0.0049), so VR-SAG does not dominate every metric. Table 5 provides the complementary NLL and AUC results: VR-SAG+TS improves NLL on AliCCP and AE, while SAG+PS has the lowest AdAuction NLL. These results support a dataset-dependent calibration trade-off, with no guarantee of improvement in every proper loss or ranking metric.

3.2 Analysis of Calibration-Error Metrics

We assess calibration metrics through a simple *rank-consistency* protocol: across many random train/validation/test splits and model variants, we compute each metric together with the oracle probability error (available on AdAuction via ground-truth CTRs) and measure the Spearman rank correlation between the trial-wise rankings they induce. This abstracts away the absolute scale of a metric and asks whether it orders predictors similarly to oracle probability error. For a fixed evaluation set, taking the square root preserves the ordering of oracle mean squared errors; neither oracle metric equals the binary-label Brier score.

As summarized in Figure 1, LCCE shows strong rank agreement with oracle probability error in the displayed setting. The additional heatmaps contain settings where equal-frequency ECE has higher correlation, so this comparison does not establish universal superiority. Appendix C gives the Monte Carlo settings, additional heatmaps and variance comparisons.

3.3 Ablation Study

We conduct ablations on AdAuction to study three hyperparameters in our method: the number of partitions, the number of groups per partition, and the weight of the intra-group variance term. Results are summarized in Figure 2.

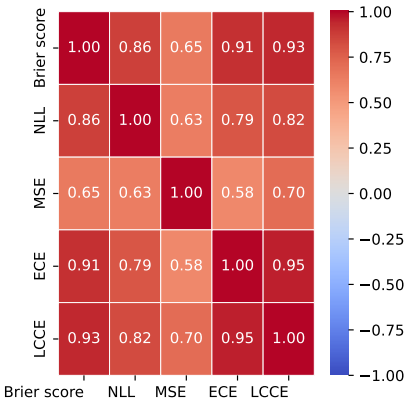
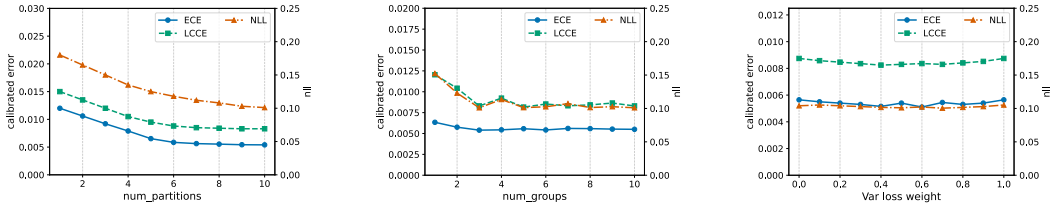


Figure 1: Spearman rank correlations between candidate calibration metrics and the historical oracle error reference over random trials. The comparison describes the displayed simulation setting.



(a) Number of partitions (b) Number of groups (c) Weight of variance loss

Figure 2: Influence of key hyperparameters in VR-SAG (means over random seeds; no error bars shown).

Figure 2 separates the number of learned partitions P, the groups per partition K, and the variance weight λ_v. The plotted means suggest benefits from increasing P over the tested range; the figure provides no error bars to quantify seed variability. In contrast, increasing the number of groups per partition yields diminishing returns. Smaller per-group sample support is one possible explanation for the limited gains as K grows. These curves do not measure group correlation or identify the mechanism behind the trend. The variance-weight curve varies modestly over the tested range. This sensitivity result does not establish the size or statistical significance of the regularizer's contribution. The matched SAG/VR-SAG comparisons in Table 1 provide the relevant empirical contrast; selecting K and λ_v requires held-out data.

We also visualize the group-wise distributions and confirm that VR-SAG yields substantially smaller within-group variance than field-based and plain SAG groupings (see Appendix H.3).

4 Related Work

Early work on click-through-rate prediction treated it as supervised learning, from logistic regression for estimating click probabilities (Richardson et al., 2007) to large-scale online systems for industrial deployment (McMahan et al., 2013). In modern ad platforms, CTR/CVR models feed real-time bidding, where the revenue-optimal bid scales with the true response probability (Zhang et al., 2014a); reinforcement-learning agents further adapt bids for budgeted campaigns (Cai et al., 2017) but still depend on calibrated upstream probabilities.

Because perfect calibration is unattainable in finite samples, practitioners approximate it by binning predictions and comparing confidence with observed outcomes (de Menezes e Silva Filho et al., 2023). Expected Calibration Error (ECE) (Guo et al., 2017a) popularized this view and motivated simple post-hoc calibrators such as histogram binning (Chaudhuri et al., 2017), isotonic regression (Menon et al., 2012; Borisov et al., 2018), and Platt scaling (Platt et al., 1999). More re-

cent "field-aware" approaches (Yang et al., 2024; Zhao et al., 2024) combine such mappings with user/item context and often reduce calibration error without degrading ranking metrics (Wei et al., 2022; Pan et al., 2020).

Field-aware calibration can be viewed as a specific instance of multi-calibration, which enforces calibration simultaneously over many (potentially overlapping) subpopulations (Hébert-Johnson et al., 2018). Beyond pre-specified partitions, a line of work learns the grouping itself using tree-based or data-driven partitioning schemes (Huang et al., 2022; Zadrozny & Elkan, 2001; Leathart et al., 2017; Durfee et al., 2022), though such groups can be cumbersome to integrate into deep recommender stacks and may optimize surrogate objectives only loosely aligned with probability calibration. Automatically discovered or latent regions (including ours) no longer coincide with simple field values such as *device* or *placement*, improving flexibility in capturing behavioral regimes but reducing direct human interpretability. Related ideas around grouping and calibration have also appeared in adjacent areas such as graph neural networks and confidence estimation for large language models (Seo et al., 2025; Zhuang et al., 2024; Detommaso et al., 2024), underscoring the broad interest in calibrated uncertainty across modern ML systems.

Positioning. VR-SAG inherits the frozen backbone and embedding-based grouping architecture from SAG (Yang et al., 2023). Its algorithmic addition is the within-group variance regularizer; PS and TS specify the component calibrators. The Brier decomposition motivates this regularizer without establishing a general calibration guarantee. LCCE supplies a separate fixed-partition evaluation. Our experiments compare classical calibrators and SAG; they do not establish superiority over recent multi-field methods (Yang et al., 2024; Zhao et al., 2024). Smooth ECE (Błasiok & Nakkiran, 2024) offers a complementary kernel-based approach to calibration measurement; we do not evaluate it here.

5 Conclusion

We presented VR-SAG, which adds within-group variance regularization to SAG for CTR calibration. The Brier decomposition separates variance from covariance and motivates the regularizer as a grouping bias. LCCE measures group-mean calibration on fixed backbone-logit clusters. The reported comparisons show improvements over the corresponding SAG variants in ECE and LCCE, with dataset-dependent NLL and AUC trade-offs.

Limitations and impact

Our evidence covers offline CTR calibration on two public datasets and one simulator. It does not establish CVR performance, resilience to temporal drift, or serving latency. Three learned groups per head represent coarse latent regions; their human interpretability and tail-group support require further analysis. Fixed logit clusters can miss within-cluster or field-specific errors. Deployment requires validation on recent held-out traffic and monitoring of calibration, ranking and group support. Changes to estimated click probabilities can affect advertiser spending and exposure allocation; offline metric gains alone do not establish revenue or fairness benefits.

Reproducibility Statement

Appendix I describes the reported split protocol, calibration settings, dataset counts and backbone source configurations. It identifies the public reference implementations for the comparative methods. We report descriptive means over 100 test-set splits; overlapping splits limit the interpretation of the original significance tests. The supplementary material includes anonymized calibration code. AliCCP and AliExpress are public datasets; we plan to release AdAuction upon acceptance.

REFERENCES

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, 2024.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, vol. 36, pp. 72071–72095, 2023.

Alexey Borisov, Julia Kiseleva, Ilya Markov, and Maarten de Rijke. Calibration: A simple way to improve click models. In *CIKM*, pp. 1503–1506, 2018. doi: 10.1145/3269206.3269260.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *WSDM*, pp. 661–670, 2017.

Sougata Chaudhuri, Abraham Bagherjeiran, and James Liu. Ranking and calibrating click-attributed purchases in performance display advertising. In *ADKDD*, pp. 7:1–7:6, 2017.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2016*, pp. 7–10, 2016.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *RecSys*, pp. 191–198, 2016.

Telmo de Menezes e Silva Filho, Hao Song, Miquel Perelló-Nieto, Raúl Santos-Rodríguez, Meelis Kull, and Peter A. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023.

Gianluca Detommaso, Martin Bertran Lopez, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. In *ICML*, 2024.

David Durfee, Aman Gupta, and Kinjal Basu. Heterogeneous calibration: A post-hoc model-agnostic framework for improved generalization. *CoRR*, vol. abs/2202.04837, 2022.

Yewen Fan, Nian Si, and Kun Zhang. Calibration matters: Tackling maximization bias in large-scale advertising recommendation systems. In *ICLR*, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, vol. 70, pp. 1321–1330, 2017a.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: A factorization-machine based neural network for CTR prediction. In *IJCAI*, pp. 1725–1731, 2017b.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, ADKDD*, pp. 5:1–5:9, 2014.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*, vol. 80, pp. 1944–1953, 2018.

Siguang Huang, Yunli Wang, Lili Mou, Huayue Zhang, Han Zhu, Chuan Yu, and Bo Zheng. MBCT: tree-based feature-aware binning for individual uncertainty calibration. In *WWW*, pp. 2236–2246, 2022.

Tim Leathart, Eibe Frank, Geoffrey Holmes, and Bernhard Pfahringer. Probability calibration trees. In *ACML*, pp. 145–160, 2017.

Kun Lin, Masoud Mansoury, Farzad Eskandanian, Milad Sabouri, and Bamshad Mobasher. Beyond static calibration: The impact of user preference dynamics on calibrated recommendation. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP*, 2024.

Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *SIGIR*, pp. 1137–1140, 2018.

H. Brendan McMahan and Omkar Muralidharan. On calibrated predictions for auction selection mechanisms. *CoRR*, vol. abs/1211.3955, 2012.

H. Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. Ad click prediction: a view from the trenches. In *KDD*, pp. 1222–1230, 2013.

Aditya Krishna Menon, Xiaoqian Jiang, Shankar Vembu, Charles Elkan, and Lucila Ohno-Machado. Predicting accurate probabilities with a ranking loss. In *ICML*, 2012.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, vol. 12, no. 4, 1973.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pp. 625–632, 2005.

Feiyang Pan, Xiang Ao, Pingzhong Tang, Min Lu, Dapeng Liu, Lei Xiao, and Qing He. Field-aware calibration: A simple and empirically strong method for reliable probabilistic predictions. In *WWW*, pp. 729–739, 2020.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW*, pp. 521–530, 2007.

Hyunjin Seo, Kyusung Seo, Joonhyung Park, and Eunho Yang. Towards precise prediction uncertainty in gnns: Refining gnns with topology-grouping strategy. In *AAAI*, pp. 20329–20337, 2025.

Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. Joint optimization of ranking and calibration with contextualized hybrid model. In *KDD*, pp. 4813–4822, 2023.

Kefan Su, Yusen Huo, Zhilin Zhang, Shuai Dou, Chuan Yu, Jian Xu, Zongqing Lu, and Bo Zheng. Auctionnet: A novel benchmark for decision-making in large-scale games. In *NeurIPS*, vol. 37, 2024.

Penghui Wei, Weimin Zhang, Ruijie Hou, Jinquan Liu, Shaoguo Liu, Liang Wang, and Bo Zheng. Posterior probability matters: Doubly-adaptive calibration for neural predictions in online advertising. In *SIGIR*, pp. 2645–2649, 2022.

Jia-Qi Yang, De-Chuan Zhan, and Le Gan. Beyond probability partitions: Calibrating neural networks with semantic aware grouping. In *NeurIPS*, vol. 36, 2023.

Shuai Yang, Hao Yang, Zhuang Zou, Linhe Xu, Shuo Yuan, and Yifan Zeng. Deep ensemble shape calibration: Multi-field post-hoc calibration in online advertising. In *KDD*, pp. 6117–6126, 2024.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pp. 694–699, 2002.

Weinan Zhang, Shuai Yuan, and Jun Wang. Optimal real-time bidding for display advertising. In *KDD*, pp. 1077–1086, 2014a.

Weinan Zhang, Shuai Yuan, and Jun Wang. Real-time bidding benchmarking with ipinyou dataset. *CoRR*, vol. abs/1407.7073, 2014b.

Yuang Zhao, Chuhan Wu, Qinglin Jia, Hong Zhu, Jia Yan, Libin Zong, Linxuan Zhang, Zhenhua Dong, and Muyu Zhang. Confidence-aware multi-field model calibration. In *CIKM*, pp. 5111–5118, 2024.

Dingyi Zhuang, Chonghe Jiang, Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. GETS: ensemble temperature scaling for calibration in graph neural networks. *CoRR*, vol. abs/2410.09570, 2024.

A Decomposing the Brier Score: From Murphy's to Semantic Grouping

Modern click-through-rate (CTR) systems emit dense, high-dimensional representations and near-continuous probability scores, yet their calibration is still assessed with tools that date back to the 1970s. To understand where miscalibration originates—and how targeted post-hoc corrections like SAG or VR-SAG can fix it—we dissect the Brier score into interpretable components. We proceed in two stages. First, we revisit Murphy's classical *Uncertainty–Reliability–Resolution* (UNC–REL–RES) decomposition for hard probability bins, clarifying its assumptions and statistical meaning. Second, we generalize the same algebra to *soft, semantics-aware regions* induced by a neural grouping head, which yields an additional variance–covariance term and exposes new levers for calibration improvement. Unless otherwise stated, expectations refer to *population* quantities; empirical ("sample") analogues follow by replacing expectations with averages over data.

A.1 Classical Murphy UNC–REL–RES Decomposition

Setup. Consider a binary event $Y \in \{0, 1\}$ and a probabilistic predictor that can output only *discrete* probability levels $p_r \in [0, 1]$, $r = 1, \ldots, R$. Let $\mathcal{I}_r$ be the index set of instances that received the level p_r, with cardinality $n_r = |\mathcal{I}_r|$, and empirical event frequency

$$o_r \;=\; \frac{1}{n_r} \sum_{i \in \mathcal{I}_r} Y_i.$$

Denote the overall empirical base rate by $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ with $n = \sum_r n_r$. The (empirical) Brier score is

$$\mathrm{BS} \;=\; \frac{1}{n} \sum_{i=1}^{n} (\, Y_i - p_{r(i)} \,)^2,$$

where $r(i)$ is the level applied to instance i.

Derivation. Add and subtract $o_{r(i)}$ inside the square:

$$\begin{aligned} \mathrm{BS} &= \frac{1}{n} \sum_{i=1}^{n} \big[(Y_i - o_{r(i)}) + (o_{r(i)} - p_{r(i)}) \big]^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - o_{r(i)})^2 + \frac{1}{n} \sum_{i=1}^{n} (o_{r(i)} - p_{r(i)})^2 + \frac{2}{n} \sum_{i=1}^{n} (Y_i - o_{r(i)})(o_{r(i)} - p_{r(i)}). \end{aligned} \tag{13}$$

Inside any fixed category r, o_r is constant, hence $\sum_{i \in \mathcal{I}_r} (Y_i - o_r)(o_r - p_r) = 0$ and the cross-term in equation 13 vanishes. Grouping the remaining terms by r yields (Murphy, 1973)

$$\mathrm{BS} \;=\; \underbrace{\bar{Y}\,(1 - \bar{Y})}_{\text{UNC}} + \underbrace{\sum_{r=1}^{R} \frac{n_r}{n}\,(p_r - o_r)^2}_{\text{REL}} - \underbrace{\sum_{r=1}^{R} \frac{n_r}{n}\,(o_r - \bar{Y})^2}_{\text{RES}}. \tag{Murphy}$$

Here, UNC is the irreducible Bernoulli variance, REL penalizes mismatch between p_r and o_r, and RES rewards partitions whose empirical frequencies o_r deviate from the global base rate $\bar{Y}$ (hence the negative sign). Because forecasts are constant within each category, the within-category variance of the scores is zero and no extra term appears. The corresponding population identity is obtained by replacing empirical averages with expectations.

A.2 Grouping Decomposition with Soft Semantic Regions

Hard vs. soft grouping. Classical reliability diagrams assume a *hard* partition: each instance x belongs to exactly one bin G_k with indicator $\mathbf{1}_{G_k}(x) \in \{0, 1\}$. In SAG/VR-SAG, the grouping head instead assigns a *soft* membership

$$q_k(x) \;=\; \big[\mathrm{softmax}\big(z(x)^\top W + b\big)\big]_k, \qquad 0 < q_k(x) < 1, \quad \sum_{k=1}^{K} q_k(x) = 1.$$

All "conditional" quantities below are interpreted as *soft* conditionals induced by q_k. For any random variable $Z = Z(x, y)$ define the (population) soft mass $w_k := \mathbb{E}[\, q_k(x)\,]$ and the soft conditional expectation

$$\mathbb{E}[Z \mid G_k] := \frac{\mathbb{E}\big[Z\, q_k(x)\big]}{\mathbb{E}\big[q_k(x)\big]} \;=\; \frac{\mathbb{E}\big[Z\, q_k(x)\big]}{w_k}.$$

With this convention,

$$\pi_k := \mathbb{E}[Y \mid G_k] = \frac{\mathbb{E}\,[Y\, q_k(x)]}{w_k}, \qquad \mu_k := \mathbb{E}[\hat{p} \mid G_k] = \frac{\mathbb{E}\,[\hat{p}\, q_k(x)]}{w_k}. \tag{soft-stats}$$

The within–group variance and covariance are defined analogously:

$$\sigma_k^2 := \mathrm{Var}(\hat{p} \mid G_k) = \frac{\mathbb{E}\big[(\hat{p} - \mu_k)^2 q_k(x)\big]}{w_k},$$

$$\gamma_k := \mathrm{Cov}(\hat{p}, Y \mid G_k) = \frac{\mathbb{E}\big[(\hat{p} - \mu_k)(Y - \pi_k)\, q_k(x)\big]}{w_k}.$$

Setting $q_k(x) = \mathbf{1}_{G_k}(x)$ recovers the standard hard-bin formulas.

Step 1: Law of total expectation. Let $S(Y, \hat{p}) = (Y - \hat{p})^2$ denote the per-instance Brier score and $\bar{\pi} := \mathbb{E}[Y]$ the global click-through rate. Then

$$\mathbb{E}[S] \;=\; \sum_k w_k\, \mathbb{E}\big[(Y - \hat{p})^2 \mid G_k\big]. \tag{1}$$

Step 2: Expand the conditional score. Since $Y^2 = Y$ for Bernoulli labels,

$$\begin{aligned}\mathbb{E}\big[(Y - \hat{p})^2 \mid G_k\big] &= \pi_k - 2\big(\mu_k \pi_k + \gamma_k\big) + \big(\mu_k^2 + \sigma_k^2\big) \\ &= \pi_k(1 - \pi_k) + (\pi_k - \mu_k)^2 + \sigma_k^2 - 2\gamma_k. \end{aligned} \tag{2}$$

Step 3: Aggregate and isolate UNC, REL, RES, Δ. A direct calculation shows $\sum_k w_k \pi_k (1 - \pi_k) = \bar{\pi}(1 - \bar{\pi}) - \mathrm{Var}(\pi_k)$. Substituting into the result of Step 2 gives the **grouping decomposition**

$$\mathbb{E}[S] \;=\; \underbrace{\bar{\pi}\big(1 - \bar{\pi}\big)}_{\text{UNC}} + \underbrace{\sum_k w_k (\mu_k - \pi_k)^2}_{\text{REL}} - \underbrace{\mathrm{Var}(\pi_k)}_{\text{RES}} + \underbrace{\sum_k w_k (\sigma_k^2 - 2\gamma_k)}_{\Delta}.$$

Here, UNC is the irreducible Bernoulli variance at the global level, REL measures calibration error inside semantic regions, RES rewards partitions whose prevalences π_k are far apart, and Δ captures the additional variance–covariance contribution introduced by allowing predictions to vary within each region. When every region collapses to a single forecast ($\sigma_k^2 = \gamma_k = 0$), Δ vanishes and the formula reduces to Murphy's classical UNC + REL − RES decomposition.

Relation to the classical Murphy decomposition. Both Murphy's UNC–REL–RES identity and the grouping decomposition above express the same Brier score as an *uncertainty* term minus a *resolution* bonus plus a *reliability* penalty; the algebraic cores coincide. The difference lies in how the data are partitioned. Murphy's derivation assumes a *hard*, forecast-value partition ($\hat{p}$ is constant inside each cell), so the within–group variance σ_k^2 and covariance γ_k vanish, yielding only three terms. In contrast, our decomposition keeps *soft* semantic regions learned by SAG/VR-SAG, preserving a spread of predictions within each region and introducing the additional

$$\Delta = \sum_k w_k\big(\sigma_k^2 - 2\gamma_k\big),$$

which is *not* sign-definite in general and quantifies the contribution of intra-group dispersion and label–score covariance.

Sign of Δ and when it is positive. Recall $\Delta = \sum_k w_k\big(\sigma_k^2 - 2\gamma_k\big)$ with $\sigma_k^2 = \mathrm{Var}(\hat{p} \mid G_k) \geq 0$ and $\gamma_k = \mathrm{Cov}(\hat{p}, Y \mid G_k)$. In general, Δ is not sign-definite: if the within-group covariance γ_k is sufficiently positive, $(\sigma_k^2 - 2\gamma_k)$ can be negative. By Cauchy–Schwarz,

$$|\gamma_k| \leq \sigma_k \sqrt{\mathrm{Var}(Y \mid G_k)} = \sigma_k \sqrt{\pi_k(1-\pi_k)},$$

hence

$$\sigma_k^2 - 2\sigma_k\sqrt{\pi_k(1-\pi_k)} \;\leq\; \sigma_k^2 - 2\gamma_k \;\leq\; \sigma_k^2 + 2\sigma_k\sqrt{\pi_k(1-\pi_k)}.$$

Therefore, Δ can be negative when many groups exhibit small σ_k but large positive γ_k. Sparsity alone does not determine the sign. For example, if $Y \mid \hat{p} \sim \mathrm{Bernoulli}(\hat{p})$ within a group, then $\gamma_k = \sigma_k^2$ and its contribution is $-w_k\sigma_k^2$. Thus a calibrated predictor can have negative Δ. The bound motivates examining variance and covariance together; it does not justify reducing variance in isolation as a guaranteed loss improvement.

Advantages of the grouping view.

- **Model-aware slicing.** Regions are induced from the backbone embedding, aligning with latent user–ad semantics rather than arbitrary probability intervals.
- **Variance-aware diagnostics.** The extra Δ term reveals when score spread (large σ_k^2) or score–label coupling (large $|\gamma_k|$) dominates the Brier loss, motivating variance-reduction strategies such as VR-SAG.
- **Practicality.** A lightweight grouping head and $K(m{+}3)$ parameters per PS head give $O(PmK)$ inference work for P heads. Hardware profiling is necessary to establish serving cost.

B GROUPING DECOMPOSITION FOR THE NEGATIVE LOG–LIKELIHOOD

We now derive an exact grouping decomposition for the *negative log–likelihood* (cross-entropy)

$$S_{\mathrm{NLL}}(Y, \hat{p}) \;=\; -\big[Y \log \hat{p} + (1-Y)\log(1-\hat{p})\big],$$

using the same soft semantic regions G_k and per–region statistics $\pi_k = \Pr(Y{=}1 \mid G_k)$, $\mu_k = \mathbb{E}[\hat{p} \mid G_k]$, $\sigma_k^2 = \mathrm{Var}(\hat{p} \mid G_k)$, and $\gamma_k = \mathrm{Cov}(\hat{p}, Y \mid G_k)$ introduced earlier.

Step 1: Law of total expectation.

$$\mathbb{E}[S_{\mathrm{NLL}}] \;=\; \sum_k w_k\, \mathbb{E}\big[S_{\mathrm{NLL}} \mid G_k\big], \qquad w_k = \Pr(G_k). \tag{14}$$

Step 2: Exact per–group expansion with integral remainder. Fix a group G_k and write $\delta := \hat{p} - \mu_k$. Consider the convex function $\phi_Y(p) = -Y\log p - (1-Y)\log(1-p)$ on $p \in (0,1)$. By Taylor's theorem with the *integral form of the remainder*, for any $p = \mu_k + \delta$,

$$\phi_Y(p) \;=\; \phi_Y(\mu_k) + \phi_Y'(\mu_k)\,\delta + \int_0^1 (1-t)\,\phi_Y''\big(\mu_k + t\delta\big)\,\delta^2\,dt,$$

where $\phi'_Y(\mu) = -\frac{Y}{\mu} + \frac{1-Y}{1-\mu}$ and $\phi''_Y(\xi) = \frac{Y}{\xi^2} + \frac{1-Y}{(1-\xi)^2} \geq 0$. Taking the conditional expectation given G_k yields

$$\mathbb{E}\big[S_{\mathrm{NLL}} \mid G_k\big] = \underbrace{\mathbb{E}\big[\phi_Y(\mu_k) \mid G_k\big]}_{\text{constant at } \mu_k} + \underbrace{\mathbb{E}\big[\phi'_Y(\mu_k)\,\delta \mid G_k\big]}_{\text{linear (covariance) term}} + \underbrace{\mathbb{E}\Big[\int_0^1 (1-t)\,\phi''_Y\big(\mu_k + t\delta\big)\,\delta^2\,dt \;\Big|\; G_k\Big]}_{=:\,R_k\,\geq 0}.$$

The constant term simplifies to the cross-entropy between Bernoulli(π_k) and Bernoulli(μ_k),

$$\mathbb{E}\big[\phi_Y(\mu_k) \mid G_k\big] = -\pi_k \log \mu_k - (1-\pi_k)\log(1-\mu_k) = H(\pi_k) + \mathrm{KL}\big(\pi_k \| \mu_k\big),$$

where $H(p) = -p\log p - (1-p)\log(1-p)$. For the linear term, using $\mathbb{E}[\delta \mid G_k] = 0$ and $\mathbb{E}[Y\delta \mid G_k] = \gamma_k$ gives

$$\mathbb{E}\big[\phi'_Y(\mu_k)\,\delta \mid G_k\big] = -\frac{\gamma_k}{\mu_k} + \frac{1}{1-\mu_k}\,\mathbb{E}\big[(1-Y)\delta \mid G_k\big] = -\frac{\gamma_k}{\mu_k} - \frac{\gamma_k}{1-\mu_k} = -\frac{\gamma_k}{\mu_k(1-\mu_k)}.$$

Hence, for each group,

$$\mathbb{E}\big[S_{\mathrm{NLL}} \mid G_k\big] = H(\pi_k) + \mathrm{KL}\big(\pi_k \| \mu_k\big) - \frac{\gamma_k}{\mu_k(1-\mu_k)} + R_k, \qquad R_k \geq 0. \tag{15}$$

Step 3: Isolate UNC, REL, RES, and the heterogeneity block. Let $\bar{\pi} = \sum_k w_k \pi_k = \mathbb{E}[Y]$ be the global prevalence. Using concavity of entropy, $H(\bar{\pi}) - \sum_k w_k H(\pi_k) \geq 0$, combine equation 14 and equation 15 to obtain

$$\mathbb{E}[S_{\mathrm{NLL}}] = \underbrace{H(\bar{\pi})}_{\mathrm{UNC}} + \underbrace{\sum_k w_k\,\mathrm{KL}\big(\pi_k \| \mu_k\big)}_{\mathrm{REL}} - \underbrace{\Big(H(\bar{\pi}) - \sum_k w_k H(\pi_k)\Big)}_{\mathrm{RES}} + \underbrace{\sum_k w_k\,R_k}_{\text{curvature–weighted variance } (\geq 0)} - \underbrace{\sum_k w_k\,\frac{\gamma_k}{\mu_k(1-\mu_k)}}_{\text{covariance correction}}.$$

This is an *exact* decomposition: the first three blocks match the classical $\mathrm{UNC} + \mathrm{REL} - \mathrm{RES}$ structure, while the last two terms together play the role of a heterogeneity component. When scores are constant within each region ($\sigma_k^2 = \gamma_k = 0$), we have $R_k = 0$ and the covariance term vanishes, recovering the Murphy-style three-term form. The curvature factor $1/\{\mu_k(1-\mu_k)\}$ shows that the NLL is more sensitive to within-group dispersion and score–label coupling when μ_k is near 0 or 1.

C THE STATISTICAL COMPARISON BETWEEN CALIBRATION METRICS

This section provides a comprehensive statistical analysis of calibration metrics, supplementing the main text's findings with detailed evaluations of metric consistency

These diagnostics describe the simulation settings and do not prove population consistency. Historical figures retain their original labels. We refer to their "Brier score" axis as the historical oracle error reference; its precise aggregation and correspondence to the archived routine require source verification. We do not assume that mean errors, error differences, or standard deviations preserve their rankings under a square-root transform. The binary-label Brier score in the theoretical decomposition remains $\mathbb{E}[(Y-r)^2]$.

C.1 RANK CORRELATION AND VARIANCE CONSISTENCY

To assess the consistency of calibration metrics, we extended the Spearman correlation analysis from the main text to include additional metrics and conducted 5000 Monte Carlo simulations across varying data distributions. Additional metrics are listed below:

Sufficient-information metrics. These metrics quantify the direct discrepancy between predicted probabilities $\hat{p}$ and true click rates π, and only when ground-truth CTR is available.

- **Oracle MAE** (historical label: Brier$^+$):

$$\text{Brier}^+ = \frac{1}{N}\sum_{i=1}^{N}|\pi_i - \hat{p}_i|$$

- **KL divergence**:

$$\text{KL}(\pi\|\hat{p}) = \frac{1}{N}\sum_{i=1}^{N}\left[\pi_i \log\frac{\pi_i}{\hat{p}_i} + (1-\pi_i)\log\frac{1-\pi_i}{1-\hat{p}_i}\right]$$

- **Generalized KL divergence**:

$$D_{\text{GKL}}(\pi\|\hat{p}) = \sum_{i=1}^{N}\left[\pi_i \log\frac{\pi_i}{\hat{p}_i} - \pi_i + \hat{p}_i\right]$$

- **Chebyshev distance**:

$$\max_i |\pi_i - \hat{p}_i|$$

- **Pearson correlation**:

$$\rho(\pi,\hat{p}) = \frac{\sum_{i=1}^{N}(\pi_i-\bar{\pi})(\hat{p}_i-\bar{\hat{p}})}{\sqrt{\sum_{i=1}^{N}(\pi_i-\bar{\pi})^2\sum_{i=1}^{N}(\hat{p}_i-\bar{\hat{p}})^2}}$$

- **Spearman correlation**:

$$\rho_s(\pi,\hat{p}) = \frac{\sum_{i=1}^{N}(r_i-\bar{r})(q_i-\bar{q})}{\sqrt{\sum_{i=1}^{N}(r_i-\bar{r})^2\sum_{i=1}^{N}(q_i-\bar{q})^2}}$$

Insufficient-information metrics These metrics quantify the direct discrepancy between predicted probabilities $\hat{p}$ and labels y_i, and only when ground-truth CTR is unavailable.

- **MSE**:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$$

- **MAE**:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{p}_i|$$

- **ECE$^+$**: G_k denotes equal-frequency bins

$$\text{ECE}^+ = \sum_{k=1}^{K}\frac{|G_k|}{N}\left|\frac{1}{|G_k|}\sum_{i\in G_k}\hat{p}_i - \frac{1}{|G_k|}\sum_{i\in G_k}y_i\right|$$

- **Absolute group-mean gap, random groups** (historical plot label: LCCE rand group):

$$\text{Gap}_{1,\text{rand}} = \sum_{k=1}^{K} w_k\,|\mu_k - \pi_k|\,, \quad w_k = \frac{|G_k|}{N}$$

- **Absolute group-mean gap, field groups** (historical plot label: LCCE field group):

$$\text{Gap}_{1,\text{field}} = \sum_{k=1}^{K} w_k\,|\mu_k - \pi_k|$$

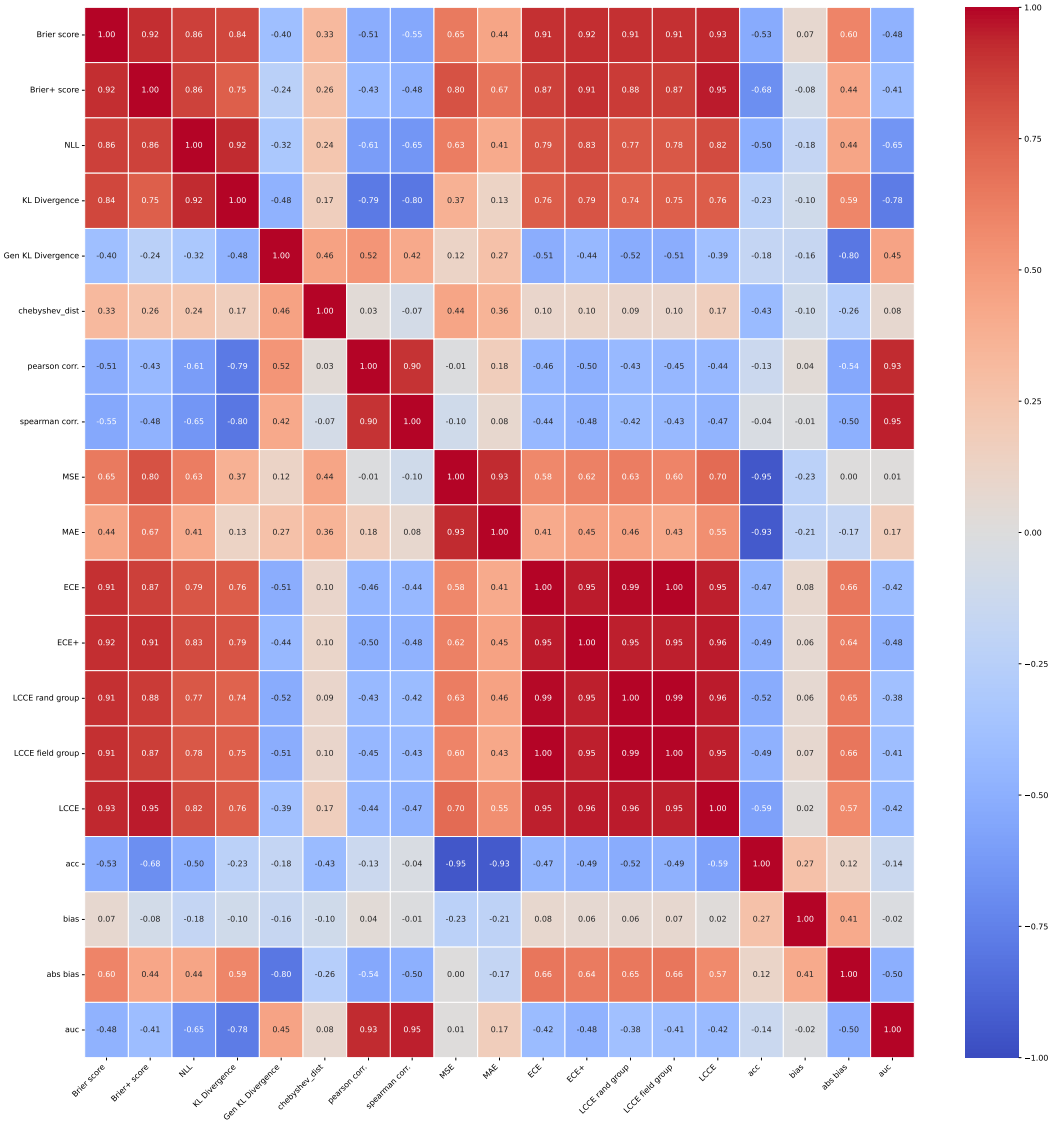


Figure 3: Spearman correlation heatmap of calibration metrics. Warmer colors indicate higher correlation, relative to the historical oracle error reference.

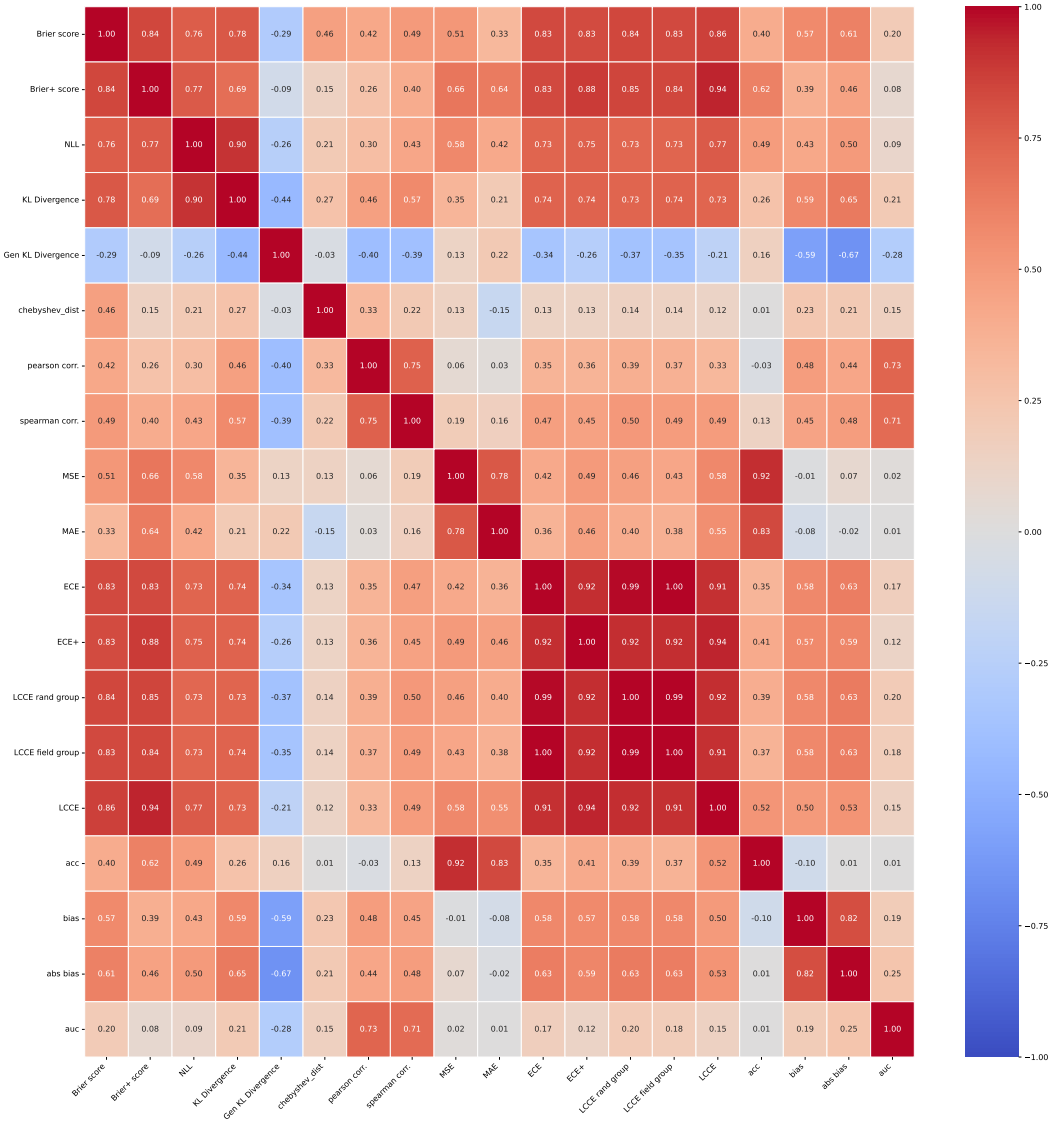


Figure 4: Spearman correlation heatmap of the standard deviations of calibration metrics and the historical oracle error reference. This is a separate descriptive comparison of variation across trials, not a consequence of rank correlation between metric means.

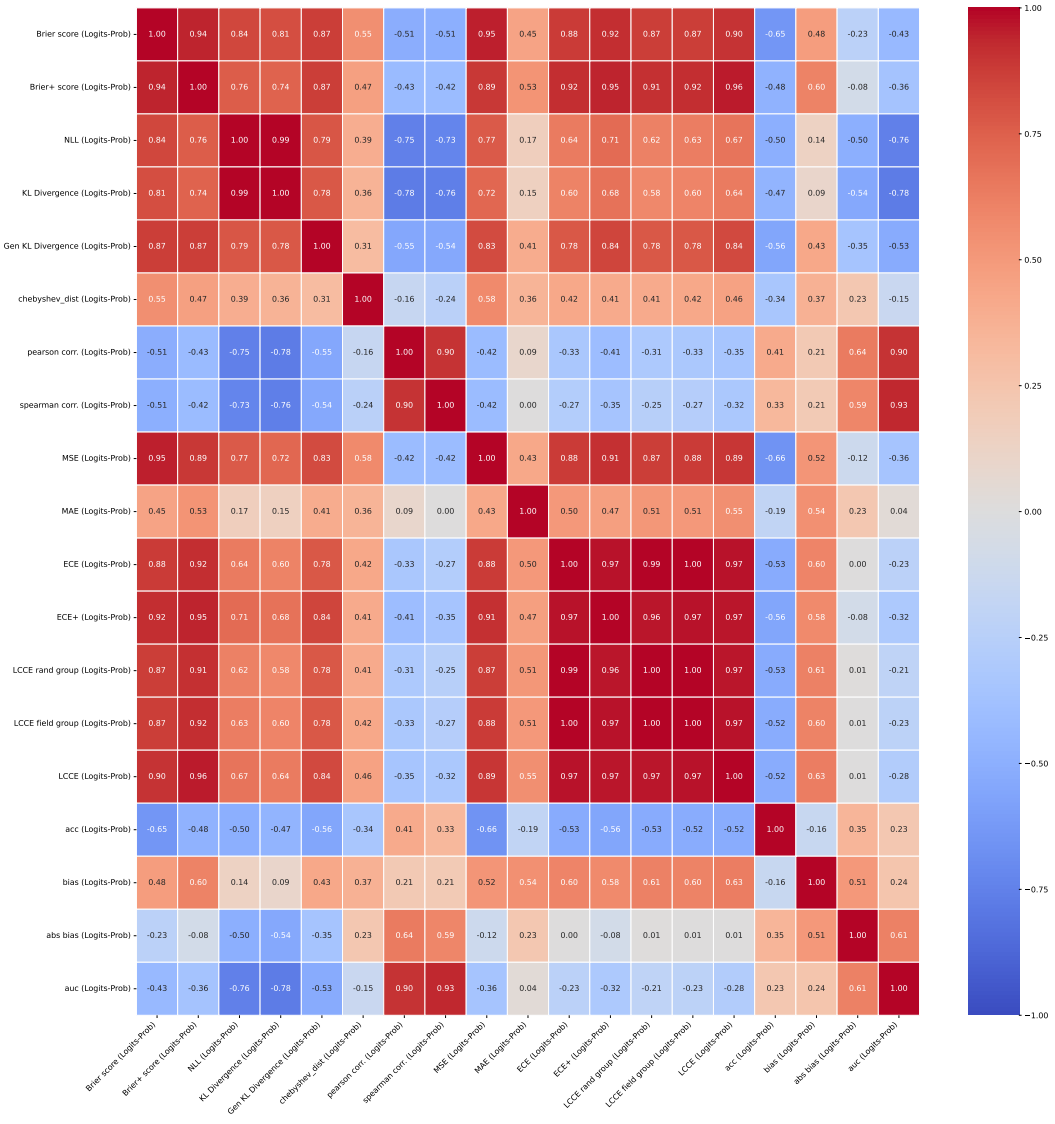


Figure 5: Spearman correlation heatmap of metric improvements (calibrated - uncalibrated). The relative correlations vary across metric pairs; this comparison does not establish a uniformly best calibration metric.

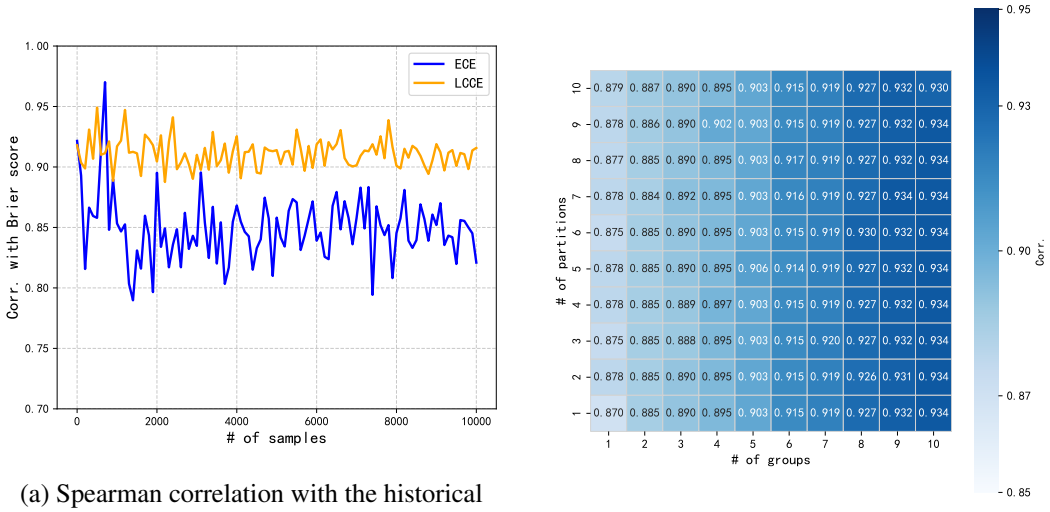
- **Bias and absolute Bias**:

$$\text{bias} = \frac{\bar{\hat{p}}}{\bar{y} + \epsilon} - 1, \quad \text{abs_bias} = \sum_k w_k \left| \frac{\bar{\hat{p}}_k}{\bar{y}_k + \epsilon} - 1 \right|, \quad \epsilon = 10^{-8}$$

These absolute-gap diagnostics differ from the squared-gap LCCE in Eq. equation 12. Figures 3, 4 and 5 present the rank correlation and variance consistency of metrics.

Figures 3–5 compare rank agreement under different simulation summaries. Strong correlation in one summary does not establish consistency or lower sampling variance in general. In the Logits-Prob improvement column of Figure 5, LCCE has correlation 0.90 with the plotted Brier reference, whereas equal-frequency ECE has 0.92. The advantage therefore depends on the comparison setting.

C.2 Convergence properties



(a) Spearman correlation with the historical oracle error reference across sample sizes.

(b) Spearman correlation of LCCE under different hyperparameter configurations.

Figure 6: Convergence behavior of calibration metrics with sample size (a) and LCCE stability under hyperparameter variations (b)

Figure 6(a) plots rank agreement as sample size increases. The LCCE calculation averages four clustering seeds, which addresses initialization variability; it does not remove sampling noise or within-cluster cancellation.

C.3 Hyperparameter sensitivity

Figure 6(b) describes the tested cluster counts and seed averages. Standard K_c-means keeps the requested number of clusters; it does not automatically merge nearby centroids. Increasing K_c trades finer resolution against lower sample support per cluster. We report sensitivity within the tested range, without claiming invariance to cluster count.

D Training Complexity

Because the compared calibration methods differ in both model structure and available implementations, asymptotic complexity alone is not very informative about their practical cost. Instead, we report wall–clock time for the full calibration pipeline (training on D_{val} and evaluating on D_{ho}) under a shared hardware setup. For all methods we rely on the authors' public implementations or faithful reimplementations, and run them on the same machine (NVIDIA 2060 Super GPU, Intel Core i7-9700 CPU, 8GB RAM).

Table 2: Training+testing time (minutes) for different calibration methods on each dataset.

	AdAuction	ALICCP	ALIEXP
Histogram binning	15	39	45
Isotonic regression	17	35	41
Platt scaling	17	35	44
Temperature scaling	20	37	44
SAG+PS	29	65	52
SAG+TS	28	64	53
VR-SAG+PS	32	97	57
VR-SAG+TS	31	93	59

As shown in Table 2, classical one-dimensional post-hoc calibrators (histogram binning, isotonic regression, Platt and temperature scaling) are the fastest, typically finishing within tens of minutes even on the largest dataset. Grouping-based methods incur additional cost from learning the grouping head and per-group parameters: SAG is moderately slower, and VR-SAG adds a further overhead because the variance penalty requires additional statistics. Reported VR-SAG pipeline times range from 31 to 97 minutes. These measurements do not establish the cost relative to backbone training or per-impression serving latency.

E Other metrics and analysis

In industrial online applications, the accurate assessment of model prediction discrepancies, including overestimations and underestimations, is of paramount importance. Among the various evaluation metrics, the predict click rate over click rate (pcoc) metric and the bias metric are frequently employed to gauge the calibration quality of model predictions. The bias metric, defined as

$$\mathrm{bias}(\hat{p}, y) = \frac{\bar{\hat{p}}}{\bar{y} + \epsilon} - 1, \qquad \mathrm{abs_bias}(\hat{p}, y) = \sum_k w_k \left| \frac{\bar{\hat{p}}_k}{\bar{y}_k + \epsilon} - 1 \right|, \quad \epsilon = 10^{-8}.$$

where k denotes grouping by media id (i.e., field-wise statistics). Bias directly quantifies the relative deviation between predicted and actual values, offering a concise and intuitive measure for identifying systematic overestimation or underestimations in model outputs. This metric captures the overall directional trend in expectations, though it incurs information loss—particularly due to the cancellation problem. For instance, positive and negative biases across different media groups may offset each other, masking true calibration errors.

The absolute bias metric mitigates this cancellation issue by summing weighted absolute deviations. Consider a scenario where one media group exhibits overestimation and another underestimation:their respective biases might cancel out, yielding a deceptively low overall bias. In contrast, the absolute bias captures such discrepancies by emphasizing the magnitude of deviations, ensuring that mis-calibrations in opposite directions are not overlooked. Specifically, a positive bias value indicates model overestimation, while a negative value signals underestimation, with the absolute bias providing a more robust measure of calibration accuracy across grouped fields.

In the following analysis, we incorporate these key industrial metrics to comprehensively evaluate the performance of each comparative method.

Table 3 presents the bias and absolute bias metrics of various calibration methods across three industrial datasets. For the AdAuction dataset, the VR-SAG+TS method achieves the smallest bias (0.0055), demonstrating its superiority in mitigating overall prediction deviation. In terms of absolute bias, which avoids cancellation of positive and negative errors, the values are generally higher than the bias metrics, highlighting the importance of using absolute bias to capture true calibration errors. Specifically, VR-SAG+PS obtains the minimum absolute bias (1.0742), outperforming other methods in quantifying the magnitude of prediction discrepancies without direction offset.For the ALICCP and AE datasets, the bias values of most calibration methods are controlled within the order of 10^{-3}, suggesting negligible overall deviation at first glance. However, this apparent "smallness" of bias metrics is misleading due to potential cancellation effects across different media groups.

Table 3: Comparison of calibration methods (reported mean values).

	AdAuction		ALICCP		AE	
Method	bias	abs_bias	bias	abs_bias	bias	abs_bias
Uncal	1.5866	13.2670	-0.039843	0.0398	-0.964428	0.9644
Histogram binning	0.0061	4.5772	-0.001060	0.0042	0.002417	0.0046
Isotonic regression	0.0086	4.1161	-0.001093	0.0042	0.002474	0.0048
Platt scaling	0.0074	4.2505	-0.001073	0.0043	0.002780	0.0048
Temperature scaling	0.0079	4.2504	-0.001071	0.0042	-0.018733	0.0187
SAG+PS	0.0075	1.7436	-0.000213	0.0036	0.000244	0.0034
SAG+TS	0.0088	1.8910	-0.000202	0.0035	-0.001183	0.0061
VR-SAG+PS	0.0074	1.0742	-0.000213	0.0036	0.000193	0.0027
VR-SAG+TS	0.0055	1.7331	-0.000213	0.0035	-0.000104	0.0054

The absolute bias metrics, though also low in magnitude, provide a more reliable assessment by emphasizing the cumulative deviation. For example, in the AE dataset, VR-SAG+PS achieves the smallest absolute bias (0.0027), indicating its robustness in handling grouped calibrations without masking errors through direction cancellation.These results underscore the critical role of absolute bias in industrial calibration evaluations, as it overcomes the limitation of traditional bias metrics that may obscure true calibration errors due to positive-negative cancellation. While the bias values on ALICCP and AE datasets suggest satisfactory performance, the absolute bias metrics reveal the nuanced differences in calibration quality, guiding the selection of more robust methods for practical applications.

E.1 Ranking changes

We report AUC changes derived from the rounded absolute AUC values in Table 5:

$$\mathrm{AUCDiff} = \frac{\mathrm{AUC}_{\mathrm{cal}} - \mathrm{AUC}_{\mathrm{uncal}}}{\mathrm{AUC}_{\mathrm{uncal}}}.$$

Table 4 uses a ratio, not percentage points. It replaces inconsistent relative values with arithmetic from one common reference table. We omit accuracy differences because the available summary does not include their absolute reference values. Global positive-temperature scaling preserves score ordering. Group-dependent transformations can reorder impressions even with a frozen backbone. The reported AliCCP AUC increase from 0.6130 to 0.6885 for VR-SAG+TS is therefore possible, but these summaries do not identify its cause. Histogram binning can introduce ties and change ordering between bins. We report the observed changes without asserting that calibration preserves AUC.

Table 4: Relative AUC change derived from Table 5. Values use ratios computed from rounded published means; averaging per-trial ratios would give a different statistic.

Method	AdAuction	AliCCP	AE
Uncal	0.0000	0.0000	0.0000
Histogram binning	-0.1784	-0.0047	-0.0354
Isotonic regression	-0.0168	-0.0003	-0.0344
Platt scaling	0.0000	0.0000	0.0000
Temperature scaling	0.0000	0.0000	0.0002
SAG+PS	-0.0080	0.0796	0.0422
SAG+TS	-0.0071	0.1250	0.0317
VR-SAG+PS	-0.0002	0.1251	0.0555
VR-SAG+TS	0.0011	0.1232	0.0568

F On the Possibility of Empty Groups and Our Handling

The main configuration uses $K = 3$ learned groups per head. Soft assignments have positive mass mathematically, but a group can have negligible effective support; a hard assignment can produce empty groups. The archived $K = 100$ diagnostic reports training-empty, test-empty and jointly empty fractions of 0.51889, 0.49944 and 0.49889. These fractions describe that diagnostic, not the main configuration or behavior under drift.

For an unsupported group, an identity correction ($\tau_k = 1$, $\beta_k = 0$) or a supported global calibrator provides a fallback. This is a deployment safeguard, not a demonstrated calibration guarantee. Group support should guide the choice of K; the available diagnostic does not establish that arbitrarily large K is safe.

G Additional Metrics—NLL and AUC

Table 5 incorporates the existing rebuttal NLL and AUC results. VR-SAG+TS has the lowest NLL on AliCCP and AE. On AdAuction, SAG+PS has the lowest NLL (0.3183); VR-SAG+PS and VR-SAG+TS are higher by 0.0008 and 0.0010. Thus improved grouping metrics need not imply improved NLL.

Table 5: Negative Log-Likelihood (NLL; lower is better) and AUC across datasets. Bold marks the smallest NLL or largest AUC in each column.

Method	AdAuction		AliCCP		AE	
	NLL	AUC	NLL	AUC	NLL	AUC
Uncal	0.3444	0.8212	0.1692	0.6130	0.3853	0.6302
Histogram binning	0.3207	0.6747	0.1702	0.6101	0.3933	0.6079
Isotonic regression	0.3199	0.8074	0.1696	0.6128	0.3922	0.6085
Platt scaling	0.3187	0.8212	0.1691	0.6130	0.3535	0.6302
Temperature scaling	0.3187	0.8212	0.1691	0.6130	0.3519	0.6303
SAG+PS	**0.3183**	0.8146	0.1658	0.6618	0.3588	0.6568
SAG+TS	0.3185	0.8154	0.1640	0.6896	0.3459	0.6502
VR-SAG+PS	0.3191	0.8210	0.1625	**0.6897**	0.3372	0.6652
VR-SAG+TS	0.3193	**0.8221**	**0.1621**	0.6885	**0.3297**	**0.6660**

These results motivate reporting calibration diagnostics, NLL and AUC together. Each measures a different aspect of probabilistic prediction.

H Practical Notes on LCCE and the Simulator

H.1 LCCE at scale and under distribution shift

A clustering subsample can reduce the cost of fitting K_c centroids. For s subsampled logits and T iterations, a direct implementation takes $O(TsK_c + NK_c)$ work, including assignment of N evaluation impressions. Table 6 preserves the archived sampling diagnostic. Relative to its 10^7 reference, the values at 10^4, 10^5 and 10^6 differ by approximately 4.11%, 0.79% and 0.000057%, respectively. This is one diagnostic, not a uniform sampling guarantee.

Under distribution shift, refitting evaluation centroids and updating a calibrator are separate operations. For a comparison within one window, all methods must share centroids. For comparisons across windows, fixed centroids preserve the partition while refitted centroids change what LCCE measures. We do not evaluate temporal robustness here.

H.2 Simulator ground-truth CTRs: rationale and validation

AdAuction supplies simulator-defined click probabilities, which permit probability-error comparisons unavailable from binary clicks alone. These probabilities are oracle values for the simulator, not observed ground truth for real users. The simulator models auction selection; the present exper-

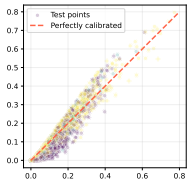
Table 6: Archived LCCE diagnostic versus clustering subsample size. Agreement with a large-sample reference measures approximation quality; a smaller value alone need not be better.

Subsample size	LCCE
10^2	0.012309877
10^3	0.021076037
10^4	0.020131670
10^5	0.019488912
10^6	0.019336675
10^7	0.019336664

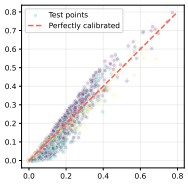
iments do not quantify its distributional fidelity to production. We plan to release AdAuction upon acceptance while keeping the internal simulator proprietary.

H.3 Verifying the Effect of the Intra-Group Variance Constraint

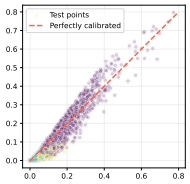
Table 1 favors VR-SAG over the corresponding SAG variants on the reported ECE and LCCE values. The following visualization compares within-group score distributions in one example.



(a) Field grouping



(b) SAG+TS



(c) VR-SAG+TS

Method	$\hat{\sigma}^2_{\text{within}}$	$\hat{\sigma}^2_{\text{between}}$
None	0.0108	0
Field	0.0106	0.0001
SAG	0.0105	0.0002
VR-SAG	0.0058	0.0050

(d) Sample variance of true probabilities

Figure 7: Grouping visualization. Backbone estimates (mean $= 0.089$) vs. true probabilities (mean $= 0.073$), with points color-coded by class labels. A 1% uniform downsampling is used for clarity. VR-SAG has the lowest within-group variance (0.0058) in this example.

The grouping visualization shows within-group homogeneity in this example. Such homogeneity is neither necessary nor sufficient for perfect calibration. Figure 7 shows lower within-group variance for VR-SAG in this example (panel (d)). In other words, even if the groups do not correspond to single human-readable fields, they form latent regions in embedding space that are explicitly more homogeneous in true CTR in this visualization.

I Reproducibility

To ensure the reproducibility of our paper, we have included all the necessary code for replicating the experiments in the supplementary materials. The code has been anonymized to maintain the anonymity of the review process. Instructions for running the code and specific implementation details for each method can be found in the README.md file and commented within the code itself.

I.1 Implementation and hyper-parameter tuning.

The implementations of the comparative methods in the paper have been modified from the corresponding open-source codes of their respective papers.

We adapted Temperature Scaling, Histogram Binning, Platt Scaling and Isotonic Regression from a public calibration implementation[3] associated with Guo et al. (Guo et al., 2017a). We adapted SAG from its public implementation[4] (Yang et al., 2023).

For some hyperparameters in the comparative methods, we follow the same setting of Yang et al. (Yang et al., 2023).

To ensure rigorous evaluation, we randomly partitioned a validation set D_{val} comprising 10% of the standard training data, alongside a hold-out set D_{ho} consisting of 10% from the standard test set for calibration purposes. For each dataset-model combination, we executed 100 test-set splits and report the mean for each method. The original analysis used paired t-tests. Overlapping test splits limit an independence-based interpretation of their significance.

Hyperparameters for comparative methods were optimized following established protocols in the literature, utilizing 5-fold cross-validation. The number of groups was fixed at 3, and the number of partitions was set to 10. Adopting a tuning strategy analogous to comparative approaches, the regularization strength was parameterized as $\lambda = 0.1$, and the group variance coefficient was specified as $\lambda_v = 0.5$

I.2 DATASETS

The offline evaluation uses AliCCP (Ma et al., 2018), AliExpress and AdAuction. Table 7 summarizes the available dataset counts.

Table 7: Dataset counts from the existing summary. CTR values are arithmetic ratios of the rounded counts, not independently measured split-level rates. Backbone AUC appears in Table 5.

Dataset	Impressions	Clicks	CTR from counts
AliCCP	42M	164K	0.00390
AliExpress	22M	574K	0.02609
AdAuction	15M	451K	0.03007

Due to the company's open-source dataset restrictions, we have only uploaded the ALICCP and ALIEXP datasets. The ALICCP backbone logits are trained and their features are extracted from open-source code [5] and the ESMM config are used. The ALIEXP backbone logits are trained and their features are extracted from open-source code [6] and the SharedBottom config are used. Upon acceptance of the paper, we will make all the AdAuction dataset publicly available together.

AI USE STATEMENT

We used generative AI for language editing, literature search, and consistency checks of mathematical exposition and reported results. This revision adds no experiments or synthetic experimental results. The authors are responsible for verifying the final manuscript.

[3] https://github.com/markus93/NN_calibration/blob/master/scripts/calibration/cal_methods.py

[4] https://github.com/ThyrixYang/group_calibration

[5] https://github.com/datawhalechina/torch-rechub/blob/main/examples/ranking/run_ali_ccp_multi_task.py

[6] https://github.com/datawhalechina/torch-rechub/blob/main/examples/ranking/run_aliexpress.py